# Unintended Bias in Misogyny Detection

Debora Nozza*
University of Milano - Bicocca
Milan, Italy
debora.nozza@unimib.it

Claudia Volpetti*
Politecnico di Milano
Milan, Italy
claudia.volpetti@polimi.it

Elisabetta Fersini
University of Milano - Bicocca
Milan, Italy
elisabetta.fersini@unimib.it

## ABSTRACT

During the last years, the phenomenon of hate against women increased exponentially especially in online environments such as microblogs. Although this alarming phenomenon has triggered many studies both from computational linguistic and machine learning points of view, less effort has been spent to analyze if those misogyny detection models are affected by an unintended bias. This can lead the models to associate unreasonably high misogynous scores to a non-misogynous text only because it contains certain terms, called *identity terms*. This work is the first attempt to address the problem of measuring and mitigating unintended bias in machine learning models trained for the misogyny detection task. We propose a novel synthetic test set that can be used as evaluation framework for measuring the unintended bias and different mitigation strategies specific for this task. Moreover, we provide a misogyny detection model that demonstrate to obtain the best classification performance in the state-of-the-art. Experimental results on recently introduced bias metrics confirm the ability of the bias mitigation treatment to reduce the unintended bias of the proposed misogyny detection model.

## CCS CONCEPTS

• **Social and professional topics** → **Hate speech**; • **Computing methodologies** → *Neural networks*.

## KEYWORDS

misogyny detection, bias measuring, bias mitigation, deep learning

**ACM Reference Format:**
Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3350546.3352512

## 1 INTRODUCTION

In the latest years, there was a growing interest in accelerating progress for women's empowerment and gender equality in our society. However, misogyny as a form of hate against them spread exponentially through the web and at very high-frequency rates, especially in online social media, where anonymity or pseudo-anonymity enables the possibility to afflict a target without being recognized or traced. This alarming phenomenon has triggered many studies related to the problem of abusive language recognition, and in particular for misogyny detection, both from computational linguistics and machine learning points of view. However, when inducing a supervised model to perform abusive language classification, it is important to focus on a particular error induced by the training data, i.e. the bias introduced in the model by a set of *identity terms* that are frequently associated to the misogynous class. For example, the term *women*, if frequently used in misogynous messages, would lead most of the supervised classification models to associate an unreasonably high misogynous score to clearly non-misogynous text, such as "You are a woman".

This behavior of recognition models is known as *unintended bias*. In particular, *"a model contains an unintended bias if it performs better for comments containing some particular identity terms than for comments containing others"* [10]. Tackling this error means being able to use those models in the real world.

In this paper, we provide a model for misogyny detection which demonstrates to obtain the best classification performance in the state-of-the-art and we address the fairness of this model by measuring and mitigating its unintended bias. In particular, to address this challenge we first propose a novel synthetic template that can be used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems. Additionally, we investigate different bias mitigation strategies, obtaining a *debiased* model that is less sensitive to identity terms as long as able to perform at the state of art of the best misogyny detection model in the literature on benchmark datasets.

Following, Section 2 provides an overview of the research works for the misogyny detection task and for the bias analysis. Then, Section 3 describes the generation process of the synthetic template test set and the investigated bias mitigation strategies. The evaluation results of the models on several recently proposed bias metrics are reported in Section 4. Finally, in Section 5, conclusions and future work are outlined.

## 2 RELATED WORK

The state-of-the-art of automatic misogyny identification in online environments is still in its infancy. A preliminary exploratory analysis of misogynous language in online social media has been presented in [17], where the authors collected and manually labeled a set of tweets as positive, negative and neutral, providing some basic statistics about the usage of some candidate misogynistic keywords. A first contribution to the problem of automatic misogyny identification has been presented in [2], where the role of different

*Both authors contributed equally to this research.


*WI '19, October 14–17, 2019, Thessaloniki, Greece*

ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352512

linguistic features and machine learning models have been investigated. More recently, thanks to the Automatic Misogyny Identification (AMI) challenges organized at IberEval [12], Evalita [13], and SemEval [4], many different approaches [3, 14, 15, 19, 20, 22] have been proposed for addressing this problem. In this context, research works commonly focus on textual feature representation studying different linguistic characteristics, ranging from pragmatical, syntactical and lexical features to higher level features derived through embedding techniques, or on the machine learning model, employing traditional or Deep Learning supervised models.

While these works focused on obtaining the most promising performance for the misogyny detection task, they do not explicitly address any study on unintended bias in their misogyny detection models. Addressing biases in text classifiers is crucial, not only because of the potentially discriminatory impact of machine learning models in real-world applications but also because bias correction can improve their robustness when used on different datasets. The research work on bias analysis can be mainly distinguished in two affiliated goals: *measuring* and *mitigating* bias.

Significant recent studies have been published on providing new metrics to quantify the presence of unintended bias in text classification models. Park et al. [21] introduce a measure of the false positive and false negative *Error Rate Equality Differences*, as a relaxation of the equalized odds fairness constraint presented in [16]. These metrics are conceived for binary labels and consequently they strictly depend on the threshold values used to separate the model output scores in two classes. In order to overcome this limitation, Dixon et al. [10] introduce a threshold agnostic metric for unintended bias called Pinned AUC, which has been proven to be inadequate in a follow-up work by the same authors [6]. Consequently, Borkan et al. [7] propose a new set of metrics differing from these early approaches because they are *(i)* threshold agnostic, *(ii)* robust to class imbalances in the dataset, and *(iii)* provide more nuanced insight into the types of bias present in the model. All the metrics cited above will be briefly introduced in Section 4.1.
On the other hand, also bias mitigation in text classification models has been significantly explored recently in the literature. Significant works [5, 10, 11, 16, 21] provide debiasing techniques ranging from debiasing word embedding to data augmentation and fine-tuning data with a larger corpus.

Our work is the first attempt to measure and mitigate unintended bias in misogyny detection models. We provide a state-of-the-art model and we test it against the most recently proposed bias metrics. Finally, we build a debiased version of our model by following the work in [10].
Moreover, since unintended bias cannot be measured on the original test set, debiasing techniques need synthetic unbiased test sets to be generated on purpose for detecting a specific bias. Previous works, such as Kiritchenko and Mohammad [18] and Park et al. [21] generated synthetic datasets for detecting gender bias. Following the identity term template method proposed in Dixon et al. [10], we also provide a novel synthetic template that can be used as the evaluation benchmark dataset for measuring unintended bias in misogyny detection task in future works and that is available online[1].

| Class | Train | Test |
|---|---|---|
| misogynous | 1,785 (45%) | 460 (46%) |
| non-misogynous | 2,215 (55%) | 540 (54%) |

**Table 1: Dataset class distribution.**

## 3 METHODOLOGY

### 3.1 Dataset

In our work, we consider the state-of-the-art corpus for misogyny detection in the English language proposed for the Automatic Misogyny Identification shared task at the Evalita 2018 evaluation campaign [12]. The corpus comprises 4,000 and 1,000 tweets for the training and test set respectively, which has been labeled by human annotators through the Figure Eight[2] crowdsourcing platform. The summary of the class distribution in the corpus is given in Table 1.

### 3.2 Identity Term Bias

In this paper, the problem of *unintended bias* is addressed by referring to the definition given by Dixon et al. [10].

*Definition 3.1.* A model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others.

This means that despite a misogyny detection model should be biased on misogynistic contents, it should not classify as misogynous tweets that explicitly refer to women or which contains women-related terms only because these are terms that usually appears in misogynistic contents. Indeed, for this study, the identity terms will be terms that can be used to refer to women, which may be unreasonably classified as misogynous with high scores.

*Identity Term List.* In order to define the list of identity terms, we take into consideration all the synonyms for "woman" by using a thesaurus[3]. The obtained list of synonymous has been then extended by including their plural form. Since some terms (e.g. *gentlewoman*) barely appear in the corpus, we decided to remove the ones with a frequency lower than 3. This choice has been made in order to study the behavior of the misogyny detection model with respect to terms that are actually seen during the training phase. The classification of instances containing identity terms that do not appear in the training set may be influenced by other factors, such as the employed sentence encoding model, exposing the unintended bias analysis to a more complex multifaceted problem which is left to future research.

*Identity Term Templates.* Since unintended bias of identity terms cannot be measured on the original test set due to class imbalance and highly different identity term contexts, *synthetic test sets* are needed to be generated on purpose.

[1] https://github.com/MIND-Lab/unintended-bias-misogyny-detection
[2] www.figure-eight.com/
[3] www.thesaurus.com

**Table 2: Template examples.**

| Template Examples | Label |
|---|---|
| <identity_term>should be protected | Non-Misogynous |
| <identity_term>should be killed | Misogynous |
| appreciate <identity_term> | Non-Misogynous |
| hit <identity_term> | Misogynous |
| amazing <identity_term> | Non-Misogynous |
| filthy <identity_term> | Misogynous |

Following previous work [10], we manually created a balanced synthetic dataset of misogynous and non-misogynous contents. We defined several templates that are filled with the previously identified identity terms and with verbs and adjectives which are divided into negative (e.g. hate, inferior) or positive (e.g. love, awesome) forms to convey hate speech or not. Table 2 reports examples of templates[4]. The generated synthetic dataset comprises 1,464 instances, of which 50% misogynous and 50% non-misogynous, where each identity term appears in the same contexts.

### 3.3 Misogyny Classification Model

With the purpose of studying the unintended bias problem in a misogyny detection model, we first build a machine learning model on the state-of-the-art misogyny corpus proposed in [12]. Then, we analyze it by measuring by using a synthetic dataset specifically designed for this task. Both datasets are introduced in the previous paragraphs. In this section, we provide details on how we designed and trained the model.

The proposed model, which we will refer to as *reference model*, is outperforming the state-of-the-art classification approaches on the misogyny corpus. We first encoded the English sentences using a novel Deep Learning Representation model, the *Universal Sentence Encoder* introduced in Cer et al. [8] built using a transformer architecture [23] and available online[5]. Once constructed the sentence embeddings, we used them as input to a single-layer neural network architecture and trained our USE_T model. To tackle the model variance, we performed 10 training runs of the same model and then we averaged the results. The model USE_T reached a 72% of mean accuracy on the test set, outperforming of two points the 70% accuracy achieved by *hateminers* team [22] ranked first to the shared task on Automatic Misogyny Identification at the Evalita 2018 evaluation campaign [12]. We implemented the model architecture using the Keras framework [9] with TensorFlow backend [1].

Since we are aware of the fact that sentence embeddings can contain biases themselves [8], we envision as future work an extended version of this study aiming to determine to what extent sentence embeddings encoded biases can affect performances in misogyny detection models.

### 3.4 Bias Mitigation Strategy

After building our reference model as described in the previous paragraph, we created four debiased versions of our USE_T model in order to mitigate its bias. This section provides further details on the bias mitigation methodologies we used.

We adopted different bias mitigation strategies motivated by the successful work by Dixon et al. [10]. The first one consists of mitigating the class imbalance of the identity terms which have the most imbalanced class distributions. After the class distribution of each identity term is computed, additional data is sampled from an external corpus and subsequently combined to the original training set in order to set the class proportions in line with the prior distribution for the overall dataset. Then, the reference model is trained on this debiased set, originating the *Debiased* model. Moreover, we also build the *Debiased_length* model, which is trained on a debiased set where the class balance is obtained also considering tweet length ranges. This permits to establish the model sensibility to the tweet length when dealing with unintended bias.

In order to confirm the benefits of the described bias mitigation procedure instead of a simple data augmentation process, we investigate the addition of randomly sampled data from the external corpus. The size of the additional random set of tweets is the same of the one computed with the aforementioned mitigation procedure. Analogously, we obtained two bias mitigated models called *Random* and *Random_length* model.

With the aim of maintaining the same language distribution of the training set for the additional data, we employed a state-of-the-art corpus for Hate Speech detection on Twitter [24] as external corpus. Tweets in the corpus have been manually annotated as sexist, racist or neither of them with almost perfect agreement. To mitigate the impact of the random sampling, both the procedures are repeated over 10 runs, originating 10 different training sets for each model.

In the following, in order to measure and evaluate our USE_T model bias, we compare it against its *Debiased*, *Debiased_length*, *Random* and *Random_length* debiased versions.

## 4 EXPERIMENTS

This section briefly describes the investigated metrics and subsequently reports their evaluation on the test set and on the generated synthetic dataset.

### 4.1 Metrics

We adopted the AUC (area under the curve) measure to evaluate the classification performance of the misogyny detection model on the test set and on the synthetic dataset. Concerning the unintended bias analysis, we computed the metrics introduced in recent state-of-the-art works [7, 10] to measure the extent of unintended bias in the model. The *Error Rate Equality Differences* measures the variation of the false positive and false negative rates between identity terms. The hypothesis motivating these metrics is that a model without unintended bias will have similar error rates across all identity terms. Since Error Rate Equality Differences measures the classification outcomes, and not the real-valued score as AUC, we applied a 0.5 threshold to discriminate between the two classes.

We decided to not investigate the Pinned AUC metric as it has been proved to suffer from several limitations [6] and that its ability to reveal unintended bias is highly impacted by a sampling procedure [10]. As suggested in [10], we investigated three separate

[4]The complete set of identity terms, verbs and adjectives is available at https://github.com/MIND-Lab/unintended-bias-misogyny-detection.
[5]https://tfhub.dev/google/universal-sentence-encoder-large/3

| Model | Test | Templates |
|---|---|---|
| USE_T | 0.7170 | 0.6339 |
| Debiased | 0.7045 | 0.6423 |
| Random | 0.7127 | 0.6396 |
| Debiased_length | 0.7003 | 0.6437 |
| Random_length | 0.7140 | 0.6376 |

**Table 3: Mean AUC on the test and synthetic templates sets.**

AUC-based metrics, recently defined in [7], which provide a more detailed view than Pinned AUC, and thus providing a more general framework for measuring unintended bias.

These metrics are calculated using the score distributions of both the whole background test data and the test set subgroup containing the identity term itself. *Subgroup AUC* (subAUC) metric provides a measure of the separability within the example from the subgroup. *Background Positive Subgroup Negative AUC* (BPSN) metric calculates AUC on the positive examples from the background and the negative examples from the subgroup. If this value is high, then it is likely that fewer negative examples from the subgroup are classified as false positives at many thresholds. *Background Negative Subgroup Positive AUC* (BNSP) metric calculates AUC on the negative examples from the background and the positive examples from the subgroup. If this value is high, then it is likely that fewer positive examples from the subgroup are classified as false negatives at many thresholds. Unfortunately, each metric provides a bias measure on a specific term exclusively. Hence, in order to combine the three per-term AUC-based metrics into one overall bias measure, we calculated their generalized mean and finally their weighted average with the overall model AUC[6], i.e. the *Weighted Bias Score*.

Additionally, two threshold agnostic metrics are studied. *Positive Average Equality Gap* (posAEG) and *Negative Average Equality Gap* (negAEG), as defined in Borkan [7], measure the separability of positive examples from the subgroup with positive examples from the background data and vice-versa. They range from -0.5 to 0.5 and their optimal value is 0. When close to the optimal value, there is no score shift from the subgroup positive examples and the background positive data since the distributions have an identical mean. The combined use of AUC-based metrics and AEGs, provide a detailed view of the types of bias present in the considered model.

## 4.2 AUC

The performance, in terms of AUC, on the test and synthetic templates sets are reported in Table 3. As a general remark, it is possible to notice that all the employed debiasing techniques have been effective on improving the mean AUC on the Identity Term Templates, while maintaining comparable performance on the test set with respect to the reference model USE_T. Comparing the results obtained with the debiasing and random treatments enables us to demonstrate that the improvements achieved by mitigating the bias are not solely due to the addition of data. The consideration of the tweet length in the bias mitigation phase has been proven to be beneficial for reducing the unintended bias.

[6]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity%2Dclassification/overview/evaluation

| Metric | False Positive Equality Difference | False Negative Equality Difference |
|---|---|---|
| USE_T | 17.49 | 20.64 |
| Debiased | 9.61 | 18.65 |
| Random | 11.44 | 26.28 |
| Debiased_length | **8.80** | **12.42** |
| Random_length | 12.18 | 26.90 |

**Table 4: Average of the Error Rate Equality Differences for each model.**

## 4.3 Error Rates

A further investigation on the analysis of unintended bias has been carried out by comparing the false positive and false negative error rates for each identity term of each model considered. It is important to mention that, with the aim of evaluating the bias, it is not important to observe the punctual values of these metrics but rather than they have similar values across all identity terms. This means that the presence of a specific identity term in a tweet is not causing an increase (or decrease) in the error rates and consequently it is not subjected to unintended bias.

Figures 1 and 2 report the false positive and false negative error rates, for each identity term, of the reference model (USE_T) and the models trained after the bias mitigation strategy considering the tweet length. Each point in the chart corresponds to the error rate of each model configuration, indeed USE_T is represented with 10 points and the bias mitigated models by 100.

By looking at false positive rates (Figure 1), it is possible to draw two different conclusions: the bias mitigation strategies, and in particular the non-random one, have (i) significantly decreased the false positive rates for each identity term and (ii) reduced the unintended bias by providing more similar values across terms.

In Figure 2, the false negative rates also demonstrate that the bias mitigation strategies are able to limit the problem of unintended bias by mitigating the differences across terms. Even if it is not essential for the bias mitigation extent, an additional consideration can be made about the absolute values of this measure, which show a different behavior from the false positive rates. In this case, the debiased models obtained higher false negative rates with a high variance among the configurations. This can be probably due to the fact that the bias mitigation strategies are specifically aimed to solve the false positive issues introducing only negative examples. Consequently, as a counter-effect, the model becomes less accurate on classifying negative examples.

## 4.4 Equality Difference Summary

In order to provide a more immediate comparison between the models, Table 4 reports the results in terms of Error Rate Equality Differences, distinguishing false positive and false negative. These results confirm the considerations made based on Figures 1 and 2, i.e. the bias mitigation strategies are reducing the unintended bias with respect to the reference model USE_T.

In particular, the improvements of the debiased model are even more evident when comparing to the model trained after the random debiasing treatment, demonstrating that the results are not

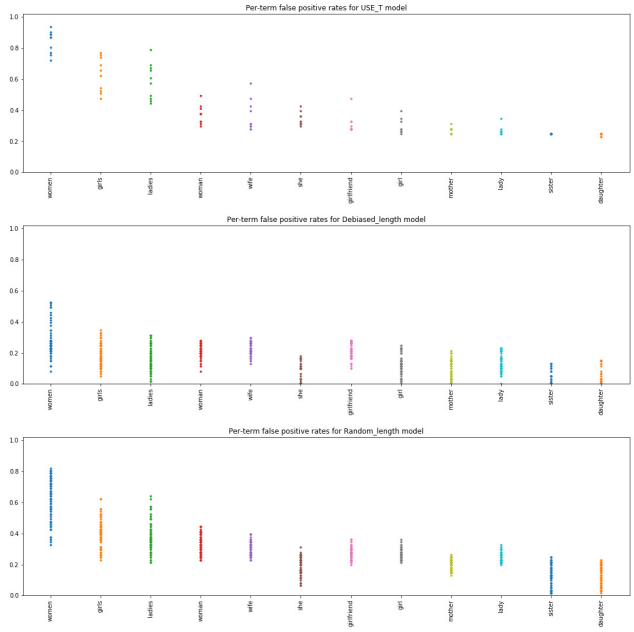


**Figure 1: False positive rates for each identity term of the reference and debiased models.**

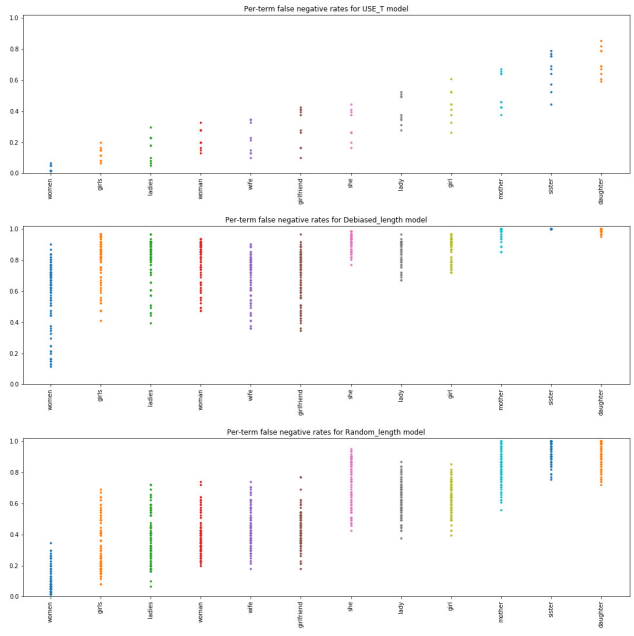


**Figure 2: False negative rates for each identity term of the reference and debiased models.**

due only to the addition of data. Moreover, it is possible to observe that the consideration of the tweet length in the bias mitigation strategy has lead to better results.

| Model | Weighted Bias Score (power mean) | Weighted Bias Score (aritmetic mean) |
|---|---|---|
| USE_T | 0.594 | 0.641 |
| Debiased | 0.593 | **0.654** |
| Random | 0.591 | 0.646 |
| Debiased_length | **0.595** | 0.652 |
| Random_length | 0.586 | 0.644 |

**Table 5: Weighted Bias Scores for each model.**

## 4.5 AUC-based metrics and AEGs

In Figure 3, we report the heatmaps for the full set of AUC-based metrics (subAUC, BPSN, BNSP) and the AEGs (negAEG, posAEG) metrics. All metrics are calculated for each identity term and heatmaps compare the USE_T reference model with the *Debiased_length* model, which demonstrated to be the most effective in reducing unintended biases according to previous analysis.

For the sake of a fair comparison, the heatmaps report the best results for each model across the sampling runs. By examining the results, we can observe that the debiased model shows a stable improvement of the subAUC measure across all terms, confirming a higher separability of positive and negative examples within each subgroup, if compared with the USE_T model subgroups separability. According to the types of biased taxonomy defined in Borkan [7], we can say that our reference model USE_T is likely to suffer from the so-called *wide subgroup score range with overlap* and *low group separability* types of bias. This can be explained by the evidence that (i) it underperformed on most of the subgroups resulting in a lower separability within subgroups compared to the background distribution and (ii) the subgroup scores distributions are so wide that they overlap with each other and with the opposite class background distributions. After the debiasing process has been applied to the model, both types of bias results mitigated, motivated by the fact that the per-subgroups AUCs are finally comparable to the mean AUC of the debiased model (see Table 3). Results in Figure 3 also show an increase in the BPSN measure on nine out of twelve sub-groups, resulting in a reduction of False Positives for the relative identity terms. A similar improvement is reported for the BNSP measure, demonstrating a reduction of the False Negatives for those subgroups that report a higher value for the metrics. Results in terms of AEGs report slight shifts of most of the subgroup distributions caused by the attempt of debiasing, but they are never reduced to their optimal value 0.

Table 5 reports the Weighted Bias Score[7], a summary metric able to combine the overall AUC with the three AUC-based metrics (subAUC, BPSN, BNSP). Debiased models outperform both the random models and the USE_T reference model, demonstrating the ability to reduce the unintended biases without losing in overall performances. A power mean (with p=-5 as suggested by authors metric) and an arithmetic mean are applied and both variants of the Weighted Bias Score results in a higher value for the debiased models with respect to the other models.

[7]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity%2Dclassification/overview/evaluation

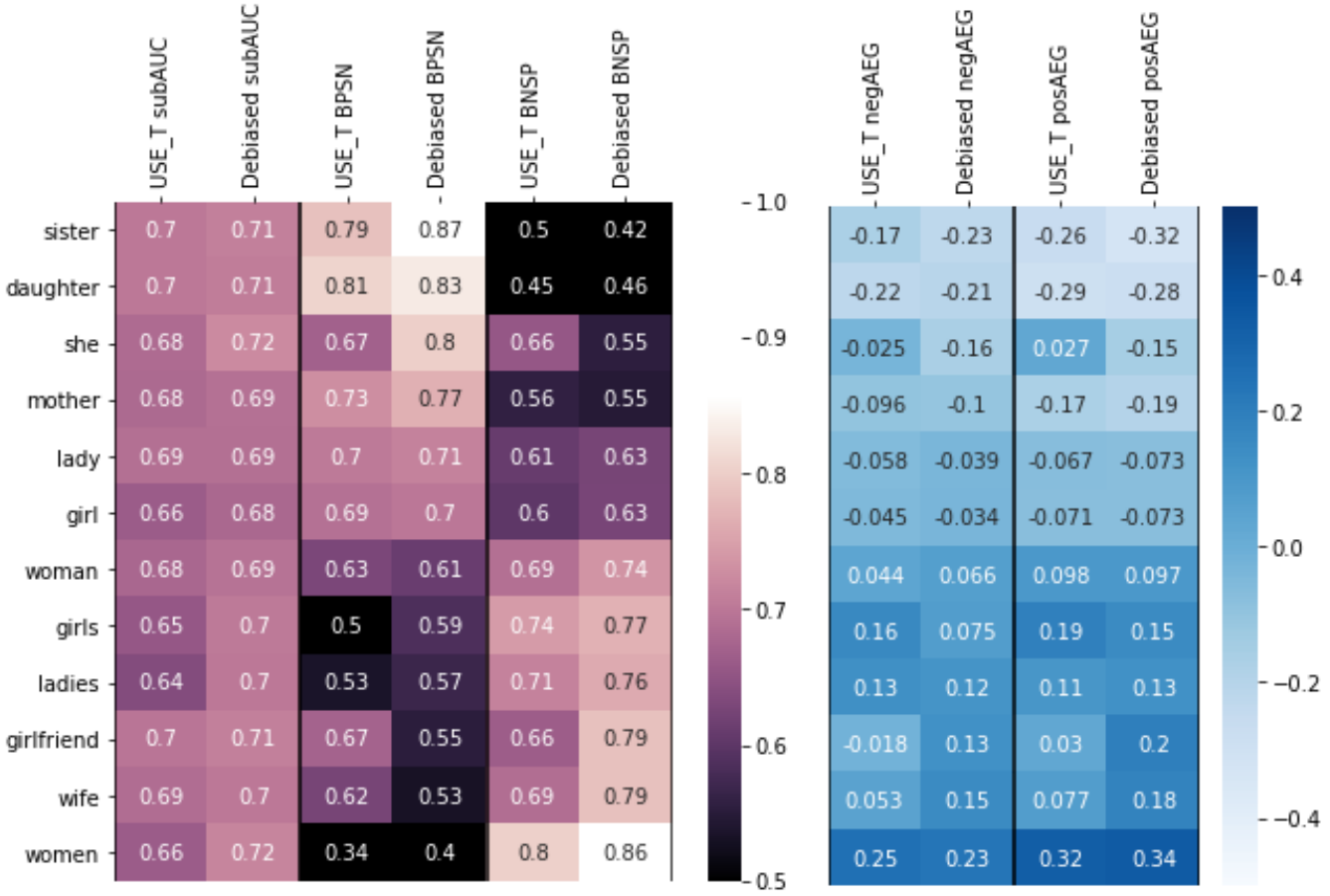


**Figure 3: Comparison between USE_T and Debiased model on the synthetic dataset.**

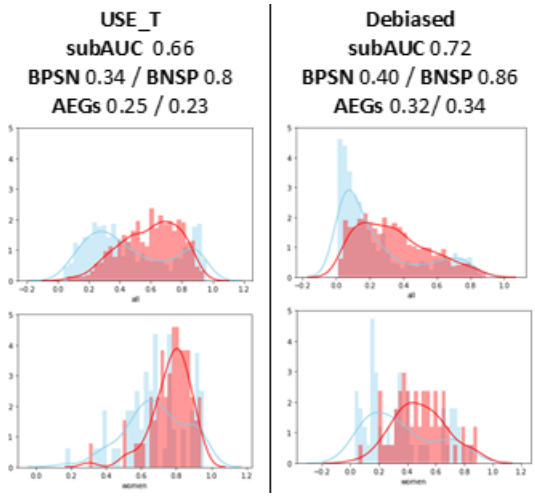


**Figure 4: Bias reduction on "women".**

Finally, Figure 4 provides an example of the debiasing method impact in reducing unintended biases for one of the most frequent identity terms in our dataset: "women". Plots reported in Figure 4 aim at graphically displaying that the subgroup separability of positive and negative examples for the debiased model is higher then the case of the reference USE_T model. This is demonstrated indeed by the increase in the subAUC value up to 0.72. This reflects on smaller numbers of False Positives and False Negatives misclassified examples. BPSN and BNSP improvements demonstrate the decrease of respectively the overlapping of negative subgroup samples with the positive background and vice-versa. Both AEGs are positives, corresponding to right-shifts of both the score distributions of the subgroup.

## 5 CONCLUSIONS AND FUTURE WORK

This paper presents the first attempt to address the problem of measuring and mitigating unintended bias in machine learning models trained for the misogyny detection task. We proposed a state-of-the-art model for misogyny detection, based on a transformer architecture, and we studied its unintended bias with some of the most recent metrics in literature.

We investigated different bias mitigation strategies, obtaining a debiased version of the proposed model that is less sensitive to identity terms as long as able to perform at the state of art of the

best misogyny detection model in the literature on benchmark datasets. The bias mitigation strategies have significantly decreased the false positive and false negative rates for each identity term and consequently reduced the unintended bias by providing more similar values across terms. The debiased model showed a stable improvement in separability of positive and negative examples within each subgroup, if compared with the reference model subgroups. Additionally, we first propose a novel synthetic template set that can be used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems.

As future work, we envision an extended version of this study aiming to determine to what extent sentence embeddings encoded biases can affect performances in misogyny detection models. The idea is to analyze and compare the impact on performances and biases of machine learning models based on pre-trained embeddings against a baseline where embeddings are trained during the learning phase.

## REFERENCES


[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018) (Lecture Notes in Computer Science)*, Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane (Eds.), Vol. 10859. Springer, 57–64.

[3] Angelo Basile and Chiara Rubagotti. 2018. CrotoneMilano for AMI at Evalita2018. A Performant, Cross-lingual Misogyny Detection System.. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

[4] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics, 54–63.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *arXiv preprint arXiv:1707.00075* (jun 2017). arXiv:1707.00075

[6] Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Limitations of Pinned AUC for Measuring Unintended Bias. *arXiv preprint arXiv:1903.02088* (2019).

[7] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion of The 2019 World Wide Web Conference (WWW 2019)*. ACM.

[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2018)*. Association for Computational Linguistics, 169–174.

[9] François Chollet et al. 2015. Keras. https://keras.io.

[10] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 67–73.

[11] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 11–21.

[12] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

[13] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*. CEUR-WS.org.

[14] Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gómez. 2018. Exploration of Misogyny in Spanish and English tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Vol. 2150. CEUR-WS.org, 260–267.

[15] Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Díaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Pérez, and Olatz Perez-de-Viñaspre. 2018. Automatic Misogyny Identification Using Neural Networks. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*. CEUR-WS.org.

[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[17] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The Problem of identifying Misogynist Language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*. ACM Press, 333–335.

[18] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM@NAACL-HLT 2018)*. Association for Computational Linguistics, 43–53.

[19] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Vol. 2150. CEUR-WS.org, 234–241.

[20] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

[21] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 2799–2804.

[22] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting Hate speech against Women. *arXiv preprint arXiv:1812.06700* (2018).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS 2017)*. 6000–6010.

[24] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Student Research Workshop, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (SRW@HLT-NAACL 2016)*. Association for Computational Linguistics, 88–93.